A statistical approach for separability of classes

D.A. Zighed, Stéphane Lallich, Fabrice Muhlenbach

A statistical approach to class separability

Djamel A. Zighed, Stéphane Lallich & Fabrice Muhlenbach
ERIC Laboratory – University of Lyon 2
5, avenue Pierre Mendès-France
F 69676 BRON Cedex – FRANCE
{zighed, lallich, fmuhlenb}@univ-lyon2.fr

Abstract: We propose a new statistical approach for characterizing the class separability degree in $\mathbb{R}^p$. This approach is based on a nonparametric statistic called "the Cut Edge Weight". We show in this paper the principle and the experimental applications of this statistic. First, we build a geometrical connected graph like Toussaint's Relative Neighbourhood Graph on all examples of the learning set. Second, we cut all edges between two examples of a different class. Third, we compute the relative weight of these cut edges. If the relative weight of the cut edges is in the expected range of a random distribution of the labels on all the neighbourhood of the graph's vertices, then no neighbourhood-based method provides a reliable prediction model. We will say then that the classes to predict are non-separable.

Keywords: Separability, Supervised Learning, Computational Geometry.

1 Introduction

Learning methods are very often requested in the data mining domain. The learning methods aim to generate a prediction model φ from a learning sample Ω_l. The model brought about is more or less reliable. This reliability is generally evaluated with *a posteriori* test sample Ω_t. The reliability depends on the learning sample, on the underlying statistical hypothesis, and many others factors. Nevertheless, it may happen that no method exist that produce a reliable model. This can be explained by the following reasons:

- the methods are not suitable to the problem we are trying to learn, we have thus to look for another more appropriate one;
- the classes are not separable in the learning space. In this case, it is impossible to find a better learning method.

It will be very interesting to use mathematical tools that can characterize the class separability from a given learning sample. There already exist measures for learnability such as the VC-dimension provided by the statistical learning theory (Vapnik 1998). Nevertheless, VC-dimension is difficult to compute in many cases. This problem has also been studied based on a statistical approach by Rao (Rao 1972). Kruskal and Wallis have defined a nonparametric test based on an equality hypothesis of the scale parameters (Aivazian, Enukov, and Mechalkine 1986). Recently, Sebban (Sebban 1996) and Zighed (Zighed and Sebban 1999) have proposed a test based on the number of edges that connect examples of different classes in a geometrical neighbourhood.

At first, they build a multidimensional neighbourhood structure by using some particular models like the Toussaint's Relative Neighbourhood Graph (Toussaint 1980). They

calculate thereafter the number of edges that must be removed from the neighbourhood graph to obtain clusters of homogeneous points in a given class. Finally, they have established the law of the edge proportion that must be removed under the null hypothesis, denoted H_0, of a random distribution of the labels. With this law, they can say if classes are separable or not by calculating the p-value of the test –e.g., the probability of having a computed value as important as the observed value under H_0.

We propose in this paper a theoretical framework and a nonparametric statistic that takes into consideration the weight of the removed edges. We exploit the works of the spatial autocorrelation, in particular the join-counts statistic, presented by Cliff and Ord (Cliff and Ord 1986) following the works of Moran (Moran 1948), Krishna Iyer (Krishna Iyer 1949), Geary (Geary 1954) and David (David 1971). Such process has been studied in the classification domain by Lebart (Lebart 2000) who used works based on the spatial contiguity, like Geary's contiguity coefficient, to compare the local structures vs. the global structures in a k nearest neighbour graph.

2 Class Separability, Clusters and Cut Edges

2.1 Notations

Machine learning methods are intended to produce a function φ that can predict the unknown belonging class $Y(\omega)$ of an instance ω extracted from the global population Ω, by knowing its representation $X(\omega)$.

In general, this representation $X(\omega)$ is provided by an expert who establishes *a priori* a set of attributes denoted: $X_1, X_2, ..., X_p$. Let these attributes take their values in $\mathbb{R}$, $X : \omega \in \Omega \mapsto X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_p(\omega)) \in \mathbb{R}^p$.

The learning sample Ω_l and a test sample Ω_t are used to build up and to assess the model φ.

The learning ability of a method is strongly associated to its class separability degree in $X(\Omega)$. We consider that the classes will be easier to separate if they fulfill the following conditions:

- the instances of the same class appear mostly gathered in the same subgroup in the representation space;
- the number of groups is small, equals the number of the classes;
- the borders between groups are simple.

2.2 Neighborhood Graphs and Clusters

To express the proximity between examples in the representation space, we use the Relative Neighbourhood Graph (RNG) of Toussaint (Toussaint 1980) defined below.

Definition: Let V be a set of points in a real space $\mathbb{R}^p$ (with p the number of attributes). The Relative Neighbourhood Graph (RNG) of V is a graph with vertices set V, and the set of edges of the RNG of V are exactly those pairs (a, b) of points for which $d(a, b) \leq Max\left(d\left(a, c\right), d\left(b, c\right)\right) \forall c, c \neq a, b$, where $d(u, v)$ denotes the distance between two points u and v in $\mathbb{R}^p$.

This definition means that the *lune* $L_{(u,v)}$ –constituted by the intersections of hypercircles centered on u and v with range the edge (u, v)– is empty. For example, on figure 1 (a), vertices 13 and 15 are connected because there is no vertex on the *lune* $L_{(13,15)}$.

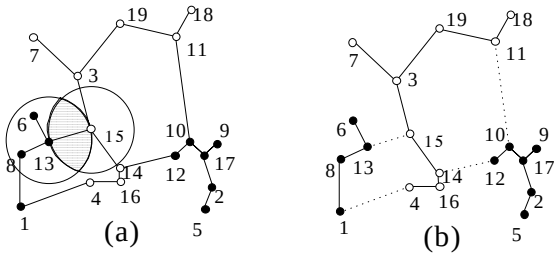

Figure 1: RNG and clusters with two classes: the black and the white points

According to Zighed and Sebban (Zighed and Sebban 1999) we introduce the concept of "cluster" to express that a set of close points have the same class. We call *cluster* a connected sub-graph of the neighbourhood graph where all vertices belong to the same class. There may be more clusters than the number of classes. To build all clusters required for characterizing the structures of the scattered data points, we proceed in two steps:

1. we generate the geometrical neighbourhood graph on the learning set;
2. we remove the edges connecting two vertices belonging to different classes, obtaining connected sub-graphs where all vertices belong to the same class.

The number of generated clusters gives a partial information on the class separability. If a number of clusters is low –at least equal to the number of classes–, the classes are well separable and we can find a learning method capable of exhibit the model that underlies the particular group structure. For example on figure 1 (b), after cutting the four edges connecting vertices of different colours (in dotted line), we obtain three clusters for the two classes. But if this number tends to increase, closely to the number of clusters that we could have in a random situation, the classes can no longer be learned due to the lack of a non random geometrical structure.

Actually, this number of clusters cannot characterize some little situations that seems intuitively different. For the same number of clusters, the situation can be very different depending on wether the clusters are easily isolated in the neighbourhood graph or not. As soon as $p > 1$, rather than studying the number of clusters, we prefer to take an interest in the edges cut for building the clusters and we will calculate the relative weight (based on the distance or the rank of the neighbourhood between two vertices) of these edges in the edge set. In our example on figure 1 (b), we have cut four edges for isolating three clusters.

3 Cut Edge Weight Statistic

In a common point between supervised classification and spatial analysis, we consider a spatial contiguity graph which plays the role of the neighbourhood graph (Cliff and Ord 1986). The vertices of this graph are coloured with k distinct colours, using for each vertex the colours corresponding to its modality. The matter is (1) to describe the link between the adjacency of two vertices and the fact they have the same colour, and (2) to test the hypothesis of non significance. This would take us to test the hypothesis of no spatial autocorrelation between the values taken by a categorical variable over spatial

units. In the case of a neighbourhood graph, this would be the results for testing the hypothesis that the class Y cannot be learned from neighbourhood-based methods.

3.1 Statistical Framework

3.1.1 Notations and Abbreviations

- Number of nodes in the graph: n
- Connection matrix: $V = (v_{ij}), i = 1, 2, ..., n; j = 1, 2, ..., n$; where $v_{ij} = 1$ if i and j are linked by an edge
- Weight matrix: $W = (w_{ij}), i = 1, 2, ..., n; j = 1, 2, ..., n$; where w_{ij} is the weight of edge (i, j). The weight w_{ij} equals: (1) $v_{ij} = 1$ (simple connection), (2) $(1 + d_{ij})^{-1}$ (weight based on the distance) or (3) ${r_i}^{-1}$ (weight based on the rank with r_j the rank of the vertex j among the neighbours of the vertex i). Let w_{i+} and w_{+j} be the sums of raw i and column j. We consider that W matrix is symmetrical (for the rank, the weights are not symmetrical, then we will use $w_{ij} = \frac{1}{2}(w_{ij} + w_{ji})$)
- Number of edges: a
- Proportion of vertices corresponding to the class y_r: π_r, $r = 1, 2, ..., k$

According to Cliff and Ord (Cliff and Ord 1986), we adopt the simplified notations below, defining some quantities used in the calculations:

Notations	Definition	Case : $W = V$
S_0	$\sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} w_{ij}$	$2a$
S_1	$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} (w_{ij} + w_{ji})^2$	$4a$
S_2	$\sum_{i=1}^{n} (w_{i+} + w_{+i})^2$	$4 \sum_{i=1}^{n} v_{i+}^2$

3.1.2 Definition of the Cut Edge Weight Statistic

In order to take into account a possible weighting of the edges, we deal with the symmetrized weights matrix W which is reduced to the connection matrix V if all the weights are equal to 1.

Edges linking two vertices of the same class (non cut edges) have to be distinguished from those linking two vertices of different classes (cut edges in order to obtain clusters).

Let us denote by I_r the sum of weights relative to edges linking two vertices of class r, and by $J_{r,s}$ the sum of weights relative to edges linking a vertex of class r and a vertex of class s. Statistics I and J are defined as it follows.

non cut edges	cut edges
$I = \sum_{r=1}^{k} I_r$	$J = \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} J_{r,s}$

In so far as I and J are connected by the relation $I + J = \frac{1}{2} S_0$, we have only to study J statistic or its normalization $\frac{J}{I+J} = \frac{2J}{S_0}$. Both give the same result after standardization. We may observe that I generalizes the test of runs in 2 dimensions and k groups (Mood 1940; Wald and Wolfowitz 1940).

3.1.3 Random Framework

Like Jain and Dubes (Jain and Dubes 1988), we consider binomial sampling in which null hypothesis is defined by:

H_0 : the vertices of the graph are labelled independently of each other, according to the same probability distribution (π_r) where π_r denotes the probability of the class $r, r = 1, 2, ..., k$.

We could consider hypergeometric sampling by adding into null hypothesis the constraint to have n_r vertices of the class $r, r = 1, 2, ...k$.

Rejecting null hypothesis means either the classes are non independently distributed or the probability distribution of the classes is not the same for the different vertices. In order to test the null hypothesis H_0 using statistic J (or I), we had first to study the distribution of these statistics under H_0.

3.2 I and J Distribution under the Null Hypothesis

To test H_0 with the statistic J, we will use two-sided tests if we are surprised by abnormally small values of J (great separability of the classes) and by abnormally large values (deterministic structuration or pattern presence). Hypothesis H_0 is rejected when J produce an extraordinary value taking into account its distribution under H_0. So, we have to establish the distribution of J under H_0 in order to calculate the p-value associated with the observed value of J as well as to calculate the critical value of J at the significance level α_0. This calculation can be done either by simulation, by permutation or by normal approximation. In the last case, we have to calculate the mean and the variance of J under H_0. According to Cliff et Ord (Cliff and Ord 1986), the proof of asymptotic normality for statistic J under binomial sampling follows from a theorem of Noether (Noether 1970): J will be asymptotically normally distributed if $S_0 - 2 \times Var(J)$ is exactly of order n^{-1}.

3.2.1 Boolean Case

The two classes defined by Y are noted 1 and 2. According to Moran (Moran 1948), $U_i = 1$ if the class of the i^{th} vertex is 1 and $U_i = 0$ if the class is 2, $i = 1, 2, ..., n$. We denote π_1 the vertex proportion of class 1 and π_2 the vertex proportion of class 2. Thus:

$$J_{1,2} = \frac{1}{2} \sum_{2} w_{ij} \left(U_i - U_j\right)^2 = \frac{1}{2} \sum_{2} w_{ij} Z_{ij}$$

where U_i are independently distributed according to Bernoulli distribution of parameter π_1, noted $B(1, \pi_1)$. It must be noticed that the variables $Z_{ij} = \left(U_i - U_j\right)^2$ are distributed according to the distribution $B(1,\, 2\pi_1\pi_2)$, but are not independent. Actually, the covariances $Cov(Z_{ij}, Z_{kl})$ are null only if the four indices are different. Otherwise, when there is a common index, one can obtain:

$$Cov(Z_{ij}, Z_{il}) = \pi_1\pi_2(1 - 4\pi_1\pi_2)$$

The table below summarizes the different results related to the statistic $J_{1,2}$:

Variable	Mean	Variance
U_i	π_1	$\pi_1\pi_2$
$Z_{ij} = \left(U_i - U_j\right)^2$	$2\pi_1\pi_2$	$2\pi_1\pi_2\left(1 - 2\pi_1\pi_2\right)$
$J_{1,2}$	$S_0\pi_1\pi_2$	$S_1\pi_1^2\pi_2^2 + S_2\pi_1\pi_2\left(\frac{1}{4} - \pi_1\pi_2\right)$
$J_{1,2}$ if $w_{ij} = v_{ij}$	$2a\pi_1\pi_2$	$4a\pi_1^2\pi_2^2 + \pi_1\pi_2\left(1 - 4\pi_1\pi_2\right)\sum_{i=1}^{n} v_{i+}^2$

The p-value of $J_{1,2}$ is calculated from standard normal distribution after centering and reducing its observed value. The critical values for $J_{1,2}$ at the significance level α_0 are:

$$J_{1,2;\alpha_{0/2}} = S_0\pi_1\pi_2 - u_{1-\alpha_{0/2}}\sqrt{S_1\pi_1^2\pi_2^2 + S_2\pi_1\pi_2\left(\frac{1}{4} - \pi_1\pi_2\right)}$$

$$J_{1,2;1-\alpha_{0/2}} = S_0\pi_1\pi_2 + u_{1-\alpha_{0/2}}\sqrt{S_1\pi_1^2\pi_2^2 + S_2\pi_1\pi_2\left(\frac{1}{4} - \pi_1\pi_2\right)}$$

3.2.2 Multiclass Case

To extend these results to the multiclass case, according to Cliff and Ord (Cliff and Ord 1986), we reason with I and J statistics already defined. These statistics are:

$I = \sum_{r=1}^{k} I_r = \frac{1}{2}\sum_2 w_{ij}T_{ij}$	$J = \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} J_{r,s} = \frac{1}{2}\sum_2 w_{ij}Z_{ij}$

where T_{ij} and Z_{ij} are random boolean variables which indicate if the vertices i and j have the same class (T_{ij}) or not (Z_{ij}).

From previous results, we easily obtain the mean of I and J:

Test statistic	Mean
$I = \sum_{r=1}^{k} I_r$	$\frac{1}{2}S_0\sum_{r=1}^{k}\pi_r^2$
$J = \sum_{r=1}^{k-1}\sum_{s=r+1}^{k} J_{r,s}$	$S_0\sum_{r=1}^{k-1}\sum_{s=r+1}^{k}\pi_r\pi_s$

Because I and J are connected by the relation $I + J = \frac{1}{2}S_0$, these two variables have the same variance, denoted $\sigma^2 = Var(I) = Var(J)$. The calculation of σ^2 is complicated due to the necessity of taking the covariances into consideration. In accordance with Cliff and Ord (Cliff and Ord 1986), we obtain the following results for binomial sampling:

$$4\sigma^2 = S_2\sum_{r=1}^{k-1}\sum_{s=r+1}^{k}\pi_r\pi_s + (2S_1 - 5S_2)\sum_{r=1}^{k-2}\sum_{s=r+1}^{k-1}\sum_{t=s+1}^{k}\pi_r\pi_s\pi_t$$
$$+4\,(S_1 - S_2)\left[\sum_{r=1}^{k-1}\sum_{s=r+1}^{k}\pi_r^2\pi_s^2 - 2\sum_{r=1}^{k-3}\sum_{s=r+1}^{k-2}\sum_{t=s+1}^{k-1}\sum_{u=t+1}^{k}\pi_r\pi_s\pi_t\pi_u\right]$$

3.3 Complexity of the Test

Different steps are into consideration: computing the matrix distance is in $O(p \times n^2)$, with n the number of examples and p the attributes, and building the neighbourhood graph in $\mathbb{R}^p$ is in $O(n^3)$. Because the number of attributes p is very small compared to the number of instances n, the test is in $O(n^3)$.

We point out that all the complete database in not needed for the test. A sample, particulary a stratified sample, can be enough to provide a good idea of the class separability in the database.

4 Experiments

4.1 Test Values on the Breiman Wave Data

We have tested the "cut edge weight" statistic on the Breiman Wave protocol described in *CART* (Breiman, Friedman, Olshen, and Stone 1984).

n	edges	clusters	$J/(I+J)$	J^S	p-value
10	9	5	0.444	-0.07	0.9408
20	25	6	0.400	-0.43	0.6668
40	56	9	0.357	-2.97	3.0E-03
60	82	11	0.354	-5.68	1.3E-08
80	115	12	0.348	-7.17	7.3E-13
100	156	12	0.301	-8.44	3.3E-17
120	187	12	0.283	-10.36	0
140	237	13	0.262	-11.59	0

Table 1: Statistical values on Breiman's Waves with different dataset sizes

In this problem, there are 3 different classes to learn with 21 predictive attributes. The wave data has been obtained for 8 different size samples : 10, 20, 40, 60, 80, 100, 120 and 1000 instances. The test values for these different samples with a relative neighbourhood graph (RNG) and a simple weight (weight = connection) are shown on table 1. In the table, n is the size of the dataset, *edges* is the the number of edges created with the RNG construction, *clusters* is the number of clusters obtained by cutting the edges between vertices of different classes, $J/(I+J)$ is the relative weight of the cut edges, J^S is the cut edge weight statistic (standardized) and *p-value* is the significance of the test (e.g., we will say that the test is significant if the p-value is lower than .05).

On table 1, we can see that the test is significant as soon as the size of the dataset is equal to 40 (p-value $< .05$): this indicates that structures are detected in the data. We can see that the number of edges grows as a linear function of the size of the dataset but the number of clusters grows as a logarithmic function of the size of the dataset.

4.2 Test Values on a Benchmark Set

The weighted edge test has been experimentally studied on 13 benchmarks from the UCI Machine Learning Repository (Blake and Merz 1998). Like the *wave* dataset, these databases have been chosen for having only numerical attributes and a symbolic class.

For each base, we build a relative neighbourhood graph (Toussaint 1980) on the n instances of the learning set. In table 2, the results show the number of instances n, the number of attributes p and the number of classes k, the information described before (the number of clusters and the number of edges) and the statistical values in three cases: when the test is done without weighting, when the edges are weighted by the inverse of the distance between the vertices, and when the edges are weighted by the inverse of the number of the rank of a vertex to the others of the graph. The last column of the "general information" is the error rate of this dataset on a 10-fold cross validation with the nearest neighbour algorithm.

The empirical evaluation of the CPU time needed for the test (distance matrix computation, graph construction, edges cut, test statistic calculation) is between a little less than 1 second for *Iris* (150 instances) and 200 seconds for *Yeast* (about 1,500 instances) on a 450 MHz PC. We present only the results obtained with a RNG graph of Toussaint (the results with a Gabriel Graph or a Minimal Spanning Tree are very close).

General information							without weighting			weighting: distance			weighting: rank		
Domain name	n	p	k	clust.	edges	error r.	J / (I + J)	J^s	p-value	J / (I + J)	J^s	p-value	J / (I + J)	J^s	p-value
Wine recognition	178	13	3	9	281	0.0389	0.093	-19.32	0	0.054	-19.40	0	0.074	-19.27	0
Breast Cancer	683	9	2	10	7562	0.0409	0.008	-25.29	0	0.003	-24.38	0	0.014	-25.02	0
Iris (Bezdek)	150	4	3	6	189	0.0533	0.090	-16.82	0	0.077	-17.01	0	0.078	-16.78	0
Iris plants	150	4	3	6	196	0.0600	0.087	-17.22	0	0.074	-17.41	0	0.076	-17.14	0
Musk "Clean1"	476	166	2	14	810	0.0650	0.167	-17.53	0	0.115	-7.69	2E-14	0.143	-18.10	0
Image seg.	210	19	7	27	268	0.1238	0.224	-29.63	0	0.141	-29.31	0	0.201	-29.88	0
Ionosphere	351	34	2	43	402	0.1397	0.137	-11.34	0	0.046	-11.07	0	0.136	-11.33	0
Waveform	1000	21	3	49	2443	0.1860	0.255	-42.75	0	0.248	-42.55	0	0.248	-42.55	0
Pima Indians	768	8	2	82	1416	0.2877	0.310	-8.74	2E-18	0.282	-9.86	0	0.305	-8.93	4E-19
Glass Ident.	214	9	6	52	275	0.3169	0.356	-12.63	0	0.315	-12.90	0	0.342	-12.93	0
Haberman	306	3	2	47	517	0.3263	0.331	-1.92	0.0544	0.321	-2.20	0.028	0.331	-1.90	0.058
Bupa	345	6	2	50	581	0.3632	0.401	-3.89	0.0001	0.385	-4.33	1E-05	0.394	-4.08	5E-05
Yeast	1484	8	10	401	2805	0.4549	0.524	-27.03	0	0.512	-27.18	0	0.509	-28.06	0

Table 2: Cut weighted edge test values on 13 benchmarks

	General information					Statistical value			Error rate						
Domain name	n	p	k	clust.	edges	J / (I + J)	J^s	p-value	1-NN	C4.5	Sipina	Perc.	MLP	N. Bayes	Mean
Breast Cancer	683	9	2	10	7562	0.008	-25.29	0	0.041	0.059	0.050	0.032	0.032	0.026	0.040
BUPA liver	345	6	2	50	581	0.401	-3.89	0.0001	0.363	0.369	0.347	0.305	0.322	0.380	0.348
Glass Ident.	214	9	6	52	275	0.356	-12.63	0	0.317	0.289	0.304	0.350	0.448	0.401	0.352
Haberman	306	3	2	47	517	0.331	-1.92	0.0544	0.326	0.310	0.294	0.241	0.275	0.284	0.288
Image seg.	210	19	7	27	268	0.224	-29.63	0	0.124	0.124	0.152	0.119	0.114	0.605	0.206
Ionosphere	351	34	2	43	402	0.137	-11.34	0	0.140	0.074	0.114	0.128	0.131	0.160	0.124
Iris (Bezdek)	150	4	3	6	189	0.090	-16.82	0	0.053	0.060	0.067	0.060	0.053	0.087	0.063
Iris plants	150	4	3	6	196	0.087	-17.22	0	0.060	0.033	0.053	0.067	0.040	0.080	0.056
Musk "Clean1"	476	166	2	14	810	0.167	-17.53	0	0.065	0.162	0.232	0.187	0.113	0.227	0.164
Pima Indians	768	8	2	82	1416	0.310	-8.74	2.4E-18	0.288	0.283	0.270	0.231	0.266	0.259	0.266
Waveform	1000	21	3	49	2443	0.255	-42.75	0	0.186	0.260	0.251	0.173	0.169	0.243	0.214
Wine recognition	178	13	3	9	281	0.093	-19.32	0	0.039	0.062	0.073	0.011	0.017	0.186	0.065
Yeast	1484	8	10	401	2805	0.524	-27.03	0	0.455	0.445	0.437	0.447	0.446	0.435	0.444
								Mean	0.189	0.195	0.203	0.181	0.187	0.259	0.202
								R^2 (J/(I+J) ; error rate)	0.933	0.934	0.937	0.912	0.877	0.528	0.979
								R^2 (J^s ; error rate)	0.076	0.020	0.019	0.036	0.063	0.005	0.026

Table 3: Error rates and statistical values of the 13 benchmarks.

4.3 Test Values and Error Rate in Machine Learning

The 13 benchmarks have been tested on the following different machine learning methods:

- instance-based learning method (the nearest neighbourhood: 1-NN (Mitchell 1997));
- decision tree (C4.5 (Quinlan 1993));
- induction graph (Sipina (Zighed, Auray, and Duru 1992));
- artificial neural networks (Perceptron (Rosenblatt 1958), Multi-Layer Perceptron with 10 neurons on one hidden layer (Mitchell 1997));
- and Naive Bayes (Mitchell 1997).

Table 3 presents the error rates obtained by these methods on a 10 cross validation with the benchmarks and the statistical values previously calculated (without weighting). The error rates for the different learning methods, and particulary the mean of these methods, are well correlated with the relative cut edge weight $(J/(I+J))$.

We can see on figure 2 the linear relation between the relative cut edge weight and the mean of the error rate for the 13 benchmarks.

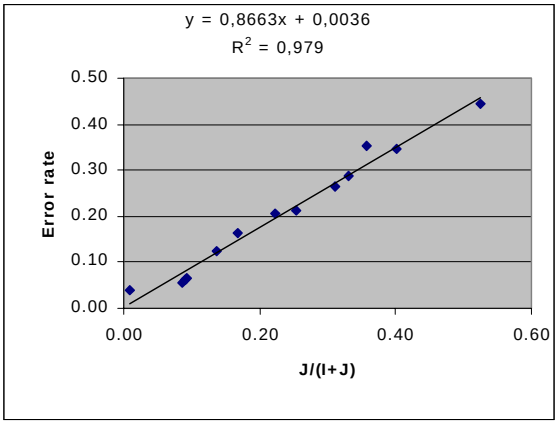

Figure 2: Relative cut edge weight and mean of the error rates.

5 Conclusion

This paper, a follow-up of Zighed and Sebban (Zighed and Sebban 1999), provides a strict framework that enables to take into consideration the weight of the edges for numerical or categorical attributes. The formalization used by Moran consists in writing the number of cut edges like a weighted sum of boolean variables. Firstly, this formalization enables exact computation for the variance of test statistic. It makes moreover possible to introduce weights in order to propose a more flexible modelling.

This framework has many applications. For example, we use it to improve classification by detecting outliers and removing and relabelling them (Lallich, Muhlenbach, and Zighed 2002). Another application is to use this framework for relevant feature selection.

The construction of the test is based on the existence of a neighbourhood graph. To build this graph, only the dissimilarity matrix is needed. This characteristic gives to our approach a very general dimension to estimate the class separability, be the instance representation known or not.

Our perspectives are to identify application fields in order to apply our method on real data. Furthermore, we plan to associate our method with visualization tools that will show the graph structure, the clusters and the contextual information about selected examples.

References

Aivazian, S., I. Enukov, and L. Mechalkine (1986). *Eléments de modélisation et traitement primaire des données*. Moscou: MIR.

Blake, C. L. and C. J. Merz (1998). UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science [http://www.ics.uci.edu/∼mlearn/MLRepository.html].

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Cliff, A. D. and J. K. Ord (1986). *Spatial processes, models and applications*. London: Pion Limited.

David, F. N. (1971). Measurement of diversity. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, USA, pp. 109–136.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician 5*, 115–145.

Jain, A. K. and R. C. Dubes (1988). *Algorithms for clustering data*. Prentice Hall.

Krishna Iyer, P. V. A. (1949). The first and second moments of some probability distribution arising from points on a lattice, and their applications. In *Biometrika*, Number 36, pp. 135–141.

Lallich, S., F. Muhlenbach, and D. A. Zighed (2002, June). Improving classification by removing or relabeling mislabeled instances. In *Foundations of Intelligent Systems, Proceedings of the 13th International Symposium on Methodologies for Intelligent Systems (ISMIS 2002), Lyon, France, June 2002*, LNAI 2366, Berlin Heidelberg, pp. 5–15. Springer-Verlag. Extended version to appear in *Journal of Intelligent Information Systems*.

Lebart, L. (2000). Data anlysis. In W. Gaul, O. Opitz, and M. Schader (Eds.), *Contiguïty analysis and classification*, Berlin, pp. 233–244. Springer.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Mood, A. (1940). The distribution theory of runs. *Ann. of Math. Statist. 11*, 367–392.

Moran, P. A. P. (1948). The interpretation of statistical maps. In *Journal of the Royal Statistical Society*, serie B, pp. 246–251.

Noether, G. E. (1970). A central limit theorem with non parametric applications. *Annals of mathematical statistics 41*, 1753–1755.

Quinlan, J. R. (1993). *C4.5: Program for Machine Learning*. San Mateo, Ca: Morgan Kaufmann.

Rao, C. R. (1972). *Linear statistical inference and its applications*. New-York: Wiley.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review 65*, 386–408.

Sebban, M. (1996). *Modèles théoriques en reconnaissance des formes et architecture hybride pour machine perceptive*. Ph. D. thesis, Université Lyon 2.

Toussaint, G. (1980). The relative neighborhood graph of a finite planar set. *Pattern recognition 12*, 261–268.

Vapnik, V. (1998). *Statistical Learning Theory*. NY: John Wiley.

Wald, A. and J. Wolfowitz (1940). On a test wether two samples are from the same population. *Ann. of Math. Statist. 11*, 147–162.

Zighed, D. A., J. P. Auray, and G. Duru (1992). *SIPINA : Méthode et logiciel*. Lacassagne.

Zighed, D. A. and M. Sebban (1999). Sélection et validation statistique de variables et de prototypes. In M. Sebban and G. Venturini (Eds.), *Apprentissage automatique*. Hermès Science.